A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Joanna Trzęsiok[*]

ON SOME SIMULATION PROCEDURES FOR COMPARING NONPARAMETRIC METHODS OF REGRESSION

Abstract. Nonparametric methods of regression form a large group of varied and rapidly growing methods. In many situations we have a problem with comparing these methods in order to select one of them to solve the regression problem. We present the simulation procedure for comparing the performance of several competing algorithms of nonparametric regression. This procedure has two stages. In the first one, the ranking of nonparametric models of regression is created. In the second stage, statistical test procedures can be used to test the significance of differences in the performances of models presented in the ranking. The procedure is applied to regression benchmark studies based on real world data.

Key words: nonparametric regression, model comparison, benchmarking experiments, hypothesis testing.

I. INTRODUCTION

The choice of the method that is the most suitable for a particular regression task is the dilemma faced by many researchers. Analyses aiming to compare and test different regression methods clearly show that it is impossible to indicate the best method which allows to build models generating the minimal mean squared errors irrespective of the data set under study (Meyer, Leisch, Hornik (2003)). The character of the data set sometimes determines the choice of a suitable method. Most of the times, however, we have a number of models to choose from and they are of equal prediction accuracy.

The paper aims to present the procedure allowing the comparison of nonparametric methods and the selection of the method most suitable for a particular regression problem. The procedure helps to create the ranking of nonparametric regression models in terms of the number of mean squared errors generated, while it takes into account the significance of the differences between the values of MSE . Due to the nature of nonparametric regression methods – their disparate mechanisms, it is impossible to analytically compare the created models. Therefore, the comparison was conducted with simulation procedures on benchmark data sets.

[*] Ph.D., Department of Mathematics, University of Economics in Katowice.

II. THE DESCRIPTION OF THE SIMULATION PROCEDURE

The simulation procedure is conducted in two stages which lead to the selection of the best solution to the regression problem. The first stage involves building many models with different – both nonparametric and linear – regression methods. Its aim is to create the ranking of the models in terms of prediction accuracy, determined based on the point estimator which is a mean squared error calculated by cross-validation (MSE_{CV}). In order to guarantee the reliability of the simulation procedure, the second stage involves the examination of the significance of the differences between the values of MSE_{CV} (calculated for the models built based on different methods). The stages of the simulation procedure are presented in detail in Table 1.

Table 1. The stages of the simulation procedure

Stage 1.	
Step 1.	Divide the D training set into 10 (approximately) equinumerous and disjoint parts.
Step 2.	Execute the following operations for each of the analysed regression methods: a) build a number of regression models for different values of the parameters of a given method; b) calculate the mean squared error by cross-validation for all models built in a); c) choose the set of parameters with the corresponding model which has the minimal MSE_{CV}; the selected model is the representative of a given method in further comparison.
Step 3.	From the training set, draw B bootstrapping samples: $L_1,\ldots,L_B$.
Step 4.	For each sample L_b (for $b=1,\ldots,B$) execute the following operations: a) divide L_b into 10 (approximately) equinumerous and disjoint parts; b) calculate the mean squared error $MSE_{CV}(M_k \mid L_b)$ by cross-validation for all regression models M_k (for $k=1,\ldots,K$) with the optimal set of values of parameters (obtained in Step 2);
Step 5.	For each model M_k (for $k=1,\ldots,K$) calculate: $MSE_{CV}(M_k)=\frac{1}{B}\sum_{b=1}^{B} MSE_{CV}(M_k \mid L_b)$.
Step 6.	Create the ranking of models M_k in terms of the values of MSE_{CV}.
Stage 2.	
Step 7.	For each pair of models M_k, M_l (for $k \neq l$ and $k,l=1,\ldots,K$) examine the significance of the differences between the values of mean squared errors $(H_0: MSE_{CV}(M_k)=MSE_{CV}(M_l))$ based on the series of values: $\{MSE_{CV}(M_k \mid L_b)\}_{b=1,\ldots,B}$ and $\{MSE_{CV}(M_l \mid L_b)\}_{b=1,\ldots,B}$. Apply the test statistic (Hothorn et al. (2005)):

	$$t = \frac{\bar{d}\sqrt{B}}{\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(d_b - \bar{d}\right)^2}},$$ which follows a t-distribution with $B-1$ degrees of freedom when $d_b = MSE_{CV}(M_k|\mathsf{L}_b) - MSE_{CV}(M_l|\mathsf{L}_b)$ and $\bar{d} = \frac{1}{B}\sum_{b=1}^{B} d_b$.
Step 8.	Correct the ranking of models from Step 6 with the results obtained in Step 7.

Source: own elaboration.

We need to emphasise that in order to assure the accuracy of testing the significance of the differences between MSE_{CV}, it is necessary to develop a uniform and clear simulation procedure which will provide the same conditions for calculations and comparisons. This, for example, means that all analysed regression models are built based on the same bootstrapping samples $\mathsf{L}_1,\ldots,\mathsf{L}_B$, drawn from a given training set. Moreover, the optimal combinations of the parameters of the models, determined at the first stage of the procedure, do not change either.

III. EMPIRICAL STUDY

The analysis was conducted on five real benchmark data sets[1]. The most important characteristics of these sets are presented in Table 2.

Table 2. The characteristics of the data sets used in the analysis

Name of the data set	Number of observations	Number of variables
Autompg	398	8
Boston	506	14
Clothing	400	13
Ozone	366	13
Star	5748	6

Source: own elaboration.

The study involved the comparison of the nonparametric regression models built using the following methods:

- projection pursuit regression (PPR) (Friedman, Stuetzle'a (1981)),
- bootstrap aggregating (BAGGING) (Breiman (1996)),

[1] Data sets used in the analysis come from libraries `Ecdat` and `mlbench` of the **R** package.

- multiple additive regression tress (MART) (Friedman (1999a), Friedman (1999b)),
- random forests (Breiman (2001)),
- multivariate adaptive polynomial spline regression (POLYMARS) (Koop-erberg et al. (1997)),
- support vector machines in regression (SVM) (Vapnik (1998)),
- neural network in regression (NNET) (cf. Bishop (1995)).

The results for the nonparametric regression models were also compared with the values of MSE_{CV}, calculated for linear model (LM).

According to the simulation procedure, the analysis was carried out in two stages and its results are presented in Tables 3-7.

In the first part of the study, we created the rankings of the regression models for each data set. The rankings were based on the mean squared errors calculated by cross-validation (this stage is illustrated with the first three columns of each of Tables 3-7).

In the second stage, we tested the significance of the differences between the values MSE_{CV}. In order to do this, we drew 100 bootstrapping samples ($B = 100$) from each training set, which means that the study used 8 (for each D set) 100-unit series of values $\left\{MSE_{CV}(M_i|\mathsf{L}_b)\right\}_{b=1,\ldots,100}$, calculated for each of the regression methods. The results of the examination of the significance of the differences between the errors MSE_{CV} resulted in a certain correction of the previously obtained rankings (presented in columns 4. and 5. of each of Tables 3-7).

Table 3. The results and rankings of regression models for the dataset *Autompg*

Stage 1.			Stage 2.	
Ranking	Methods	MSE_{CV}	Ranking	Methods
1	R. FOREST	4.04	1	R. FOREST
2	MART	5.55	2	MART
3	BAGGING	6.45	**3**	**BAGGING**
4	SVM	6.53	**3**	**SVM**
5	POLYMARS	7.45	5	POLYMARS
6	PPR	7.62	6	PPR
7	NNET	8.75	7	NNET
8	LM	11.11	8	LM

Source: own elaboration.

Table 4. The results and rankings of regression models for the dataset *Boston*

Stage 1.			Stage 2.	
Ranking	Methods	MSE_{CV}	Ranking	Methods
1	R. FOREST	5.74	1	R. FOREST
2	MART	8.21	2	MART
3	BAGGING	10.15	**3**	**BAGGING**
4	PPR	10.31	**3**	**PPR**
5	POLYMARS	11.85	***5***	***POLYMARS***
6	SVM	12.31	***5***	***SVM***
7	NNET	14.13	*6*	*NNET*
8	LM	22.70	8	LM

Source: own elaboration.

Table 5. The results and rankings of regression models for the dataset *Clothing*

Stage 1.			Stage 2.	
Ranking	Methods	MSE_{CV}	Ranking	Methods
1	PPR	$10525 \cdot 10^6$	1	PPR
2	SVM	$22417 \cdot 10^6$	2	SVM
3	MART	$38486 \cdot 10^6$	3	MART
4	R. FOREST	$47579 \cdot 10^6$	4	R. FOREST
5	BAGGING	$62471 \cdot 10^6$	5	BAGGING
6	NNET	$68114 \cdot 10^6$	6	NNET
7	LM	$82610 \cdot 10^6$	**7**	**LM**
8	POLYMARS	$94507 \cdot 10^9$	**7**	**POLYMARS**

Source: own elaboration.

Table 6. The results and rankings of regression models for the dataset *Ozone*

Stage 1.			Stage 2.	
Ranking	Methods	MSE_{CV}	Ranking	Methods
1	R. FOREST	8.93	1	R. FOREST
2	MART	9.45	2	MART
3	BAGGING	11.27	**3**	**BAGGING**
4	SVM	11.67	**3**	**SVM**
5	NNET	13.08	5	NNET
6	POLYMARS	14.59	6	POLYMARS
7	PPR	17.06	7	PPR
8	LM	19.17	8	LM

Source: own elaboration.

Table 7. The results and rankings of regression models for the dataset *Star*

Stage 1.			Stage 2.	
Ranking	Methods	MSE_{CV}	Ranking	Methods
1	R. FOREST	1 812.1	1	R. FOREST
2	MART	1 963.7	2	MART
3	PPR	1 988.3	3	PPR
4	NNET	2 037.8	4	NNET
5	BAGGING	2 041.7	5	BAGGING
6	SVM	2 052.2	6	SVM
7	POLYMARS	2 082.2	7	POLYMARS
8	LM	2 088.7	8	LM

Source: own elaboration.

The cases, where we fail to reject a null hypothesis about the insignificance of the differences between the values of MSE_{CV}, are highlighted in bold or italics in Tables 3-7. For example, for the data sets *Autompg* and *Ozon*, the values of mean squared errors calculated with BAGGING and SVM methods turned out to be insignificant, whereas the values of MSE_{CV} calculated for different regression models built on the set *Star* were significantly different in each case.

The most interesting results were obtained for the data set *Boston*. The models built with BAGGING and PPR, POLYMARS and SVM as well as SVM and NNET had insignificantly different values of the mean squared error. However, the difference between the values of MSE_{CV} for the models POLYMARS and NNET turned out to be significant.

IV. CONCLUSION

The paper discusses the simulation procedure which allows the comparison of different nonparametric regression models and the selection of the best model. The procedure is implemented in two stages. In the first stage, the ranking of regression models is created based on prediction accuracy measured with a mean squared error calculated by cross-validation (MSE_{CV}). The second stage of the analysis aims to test the significance of the differences between the obtained values of MSE_{CV} and, as a consequence, correct the rankings.

The empirical study showed that the models characterized with the best prediction accuracy were usually the models built using regression trees – most commonly the RANDOM FORESTS, but the good results were also obtained

for the MART and BAGGING models. In each of the analysed cases, the values of MSE_{CV} for the best model were significantly different from the values of MSE_{CV} calculated for the models which ranked lower.

REFERENCES

Bishop C. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.
Breiman L. (1996), Bagging Predictors, *Machine Learning*, 24, 123-140.
Breiman L. (2001), Random Forests, *Machine Learning*, 45, 5-32.
Friedman J. (1999a), *Greedy Function Approximation: a Gradient Boosting Machine*, Technical Report, Stanford University, Dept. of Statistics.
Friedman J. (1999b), *Stochastic Gradient Boosting*, Technical Report, Stanford University, Dept. of Statistics.
Friedman J., Stuetzle W. (1981), Projection Pursuit Regression, *Journal of the American Statistical Association*, 76, 817-823.
Hothorn T., Leisch F., Zeileis A., Hornik K. (2005), The Design and Analysis of Benchmark Experiments, *Journal of Computational and Graphical Statistics*, 14(3), 675-699.
Kooperberg C., Bose S., Stone C. (1997), Polychotomous Regression, *Journal of the American Statistical Association*, 92, 117-127.
Meyer D., Leisch F., Hornik K. (2003), The Support Vector Machine under Test, *Neurocomputing*, 55(1-2), 169-186.
Vapnik V. (1998), *Statistical Learning Theory*, „Adaptive and Learning Systems for Signal Processing, Communications, and Control", John Wiley & Sons, New York.

Joanna Trzęsiok

WYBRANE SYMULACYJNE PROCEDURY PORÓWNYWANIA NIEPARAMETRYCZNYCH METOD REGRESJI

W artykule przedstawiono symulacyjną procedurę badawczą pozwalającą na porównywanie różnych nieparametrycznych modeli regresji, jak i wybór najlepszego z nich. Zaproponowana procedura przebiega dwuetapowo. W pierwszym etapie tworzony jest ranking modeli regresji, pod względem dokładności predykcji, mierzonej za pomocą błędu średniokwadratowego obliczonego metodą sprawdzania krzyżowego (MSE_{CV}). Drugi etap analizy ma na celu zbadanie istotności różnic pomiędzy uzyskanymi wartościami MSE_{CV} , a tym samym skorygowanie otrzymanych rankingów. Do testowania istotności wspomnianych różnic wykorzystano nieparametryczną statystykę testującą zaproponowaną przez Hothorna. Opisaną procedurę badawczą zastosowano w badaniu empirycznym, dla zbiorów danych standardowo wykorzystywanych do analizowania własności różnych metod regresji.